Auditing large language models: a three-layered approach

Year: 2023 | Citations: 259 | Authors: Jakob Mokander, Jonas Schuett, Hannah Rose Kirk

---

Abstract

Large language models (LLMs) represent a major advance in artificial intelligence (AI) research. However, the widespread use of LLMs is also coupled with significant ethical and social challenges. Previous research has pointed towards auditing as a promising governance mechanism to help ensure that AI systems are designed and deployed in ways that are ethical, legal, and technically robust. However, existing auditing procedures fail to address the governance challenges posed by LLMs, which display emergent capabilities and are adaptable to a wide range of downstream tasks. In this article, we address that gap by outlining a novel blueprint for how to audit LLMs. Specifically, we propose a three-layered approach, whereby governance audits (of technology providers that design and disseminate LLMs), model audits (of LLMs after pre-training but prior to their release), and application audits (of applications based on LLMs) complement and inform each other. We show how audits, when conducted in a structured and coordinated manner on all three levels, can be a feasible and effective mechanism for identifying and managing some of the ethical and social risks posed by LLMs. However, it is important to remain realistic about what auditing can reasonably be expected to achieve. Therefore, we discuss the limitations not only of our three-layered approach but also of the prospect of auditing LLMs at all. Ultimately, this article seeks to expand the methodological toolkit available to technology providers and policymakers who wish to analyse and evaluate LLMs from technical, ethical, and legal perspectives.